

A Curvature Guided Composite Kernel Framework for Differential Gene Selection in Cancer Transcriptomics

Manan Gupta[†] and Arka Prava Sarkar^{*,‡}

[†]*School of Information Science, Presidency University, Bengaluru, Karnataka, India -
560119*

[‡]*Condensed Matter and Materials Physics Department, S.N. Bose National Centre for
Basic Sciences, Kolkata, West Bengal, India - 700106*

E-mail: arkaps@bose.res.in

Supporting Information

List of Tables

S1	Core Parameters Used in the proposed method-based feature selection framework	S7
S2	Classification Experiment Parameters for the Experimental Pipeline	S8
S3	Differential Expression Analysis Parameters for DESeq2, Voom-Limma, and NOISeq Pipelines	S9

List of Figures

S1	Gene set enrichment analysis and KEGG pathway enrichment analysis for DEGs selected from TCGA LUAD	S10
S2	Gene set enrichment analysis and KEGG pathway enrichment analysis for DEGs selected from TCGA HNSC	S11
S3	Gene set enrichment analysis and KEGG pathway enrichment analysis for DEGs selected from TCGA BRCA	S12
S4	Gene set enrichment analysis and KEGG pathway enrichment analysis for DEGs selected from Gene Expression Omnibus Liver Cancer	S13
S5	Gene set enrichment analysis and KEGG pathway enrichment analysis for DEGs selected from Gene Expression Omnibus Throat Cancer	S14
S6	Gene set enrichment analysis and KEGG pathway enrichment analysis for DEGs selected from Gene Expression Omnibus Pancreatic Cancer	S15

Appendix A: Pseudo-code of Algorithms

Algorithm 1 PROJECTION KERNEL

Require: Expression matrix $X_{\text{cpu}} \in \mathbb{R}^{N_{\text{samples}} \times N_{\text{genes}}}$, label vector $labels_t$, batch size $batch_{\text{genes}}$

Ensure: Kernel vector $K_c \in \mathbb{R}^{N_{\text{genes}}}$

```

1:  $(N_{\text{samples}}, N_{\text{genes}}) \leftarrow \text{shape}(X_{\text{cpu}})$ 
2: if  $|labels_t| < N_{\text{samples}}$  then
3:   raise VALUEERROR
4: else if  $|labels_t| > N_{\text{samples}}$  then
5:   truncate  $labels_t$  to the first  $N_{\text{samples}}$  entries
6: end if
7:  $K_c \leftarrow \mathbf{0} \in \mathbb{R}^{N_{\text{genes}}}$ 
8:  $normal_{\text{mask}} \leftarrow (labels_t = 0)$ 
9:  $cancer_{\text{mask}} \leftarrow (labels_t = 1)$ 
10: for  $start \leftarrow 0$  to  $N_{\text{genes}} - 1$  step  $batch_{\text{genes}}$  do
11:    $end \leftarrow \min(start + batch_{\text{genes}}, N_{\text{genes}})$ 
12:    $batch_{\text{idx}} \leftarrow start : end$ 
13:    $X_{\text{batch}} \leftarrow \text{toDevice}(X_{\text{cpu}}[:, batch_{\text{idx}}])$ 
14:    $X_{\text{normal}} \leftarrow X_{\text{batch}}[normal_{\text{mask}}]$ 
15:    $X_{\text{cancer}} \leftarrow X_{\text{batch}}[cancer_{\text{mask}}]$ 
16:    $kc_{\text{normal}} \leftarrow \sum(X_{\text{normal}}^2)$  ▷ column-wise
17:    $kc_{\text{cancer}} \leftarrow \sum(X_{\text{cancer}}^2)$  ▷ column-wise
18:    $K_c[batch_{\text{idx}}] \leftarrow kc_{\text{normal}} + kc_{\text{cancer}}$ 
19: end for
20: return  $K_c$ 

```

Algorithm 2 ESTIMATE CURVATURE

Require: Kernel vector $kernel_{\text{values}}$, neighborhood size $k_{\text{neighbors}}$

Ensure: Normalized curvature vector $curv_{\text{norm}}$

```
1:  $kc \leftarrow \text{flatten}(kernel_{\text{values}})$  on CPU
2:  $n \leftarrow |kc|$ 
3:  $(kc_{\text{sorted}}, idx_{\text{sorted}}) \leftarrow \text{SortAscending}(kc)$ 
4:  $window \leftarrow k_{\text{neighbors}}$ 
5:  $curv \leftarrow \mathbf{0} \in \mathbb{R}^n$ 
6: for  $pos \leftarrow 0$  to  $n - 1$  do
7:    $left \leftarrow \max(0, pos - window)$ 
8:    $right \leftarrow \min(n, pos + window + 1)$ 
9:    $neighbors \leftarrow kc_{\text{sorted}}[left : right]$ 
10:   $center \leftarrow kc_{\text{sorted}}[pos]$ 
11:   $dists \leftarrow |neighbors - center|$ 
12:  if  $|dists| > 1$  then
13:     $local_{\text{var}} \leftarrow \text{Var}(dists)$ 
14:  else
15:     $local_{\text{var}} \leftarrow 0$ 
16:  end if
17:   $curv[pos] \leftarrow -local_{\text{var}}$ 
18: end for
19:  $curv_{\text{unsorted}} \leftarrow \mathbf{0} \in \mathbb{R}^n$ 
20:  $curv_{\text{unsorted}}[idx_{\text{sorted}}] \leftarrow curv$ 
21:  $\mu \leftarrow \text{Mean}(curv_{\text{unsorted}})$ 
22:  $\sigma \leftarrow \text{Std}(curv_{\text{unsorted}}) + 10^{-8}$ 
23:  $curv_{\text{norm}} \leftarrow \frac{curv_{\text{unsorted}} - \mu}{\sigma}$ 
24: return  $curv_{\text{norm}}$ 
```

Algorithm 3 BUILD NEIGHBOR INDICES

Require: Kernel vector $kernel_{\text{values}}$, neighborhood size $k_{\text{neighbors}}$

Ensure: Neighbor index matrix $neighbors_{\text{unsorted}}$

```
1:  $kc \leftarrow \text{flatten}(kernel_{\text{values}})$  on CPU
2:  $n \leftarrow |kc|$ 
3:  $(kc_{\text{sorted}}, idx_{\text{sorted}}) \leftarrow \text{SortAscending}(kc)$ 
4:  $window \leftarrow \lfloor k_{\text{neighbors}}/2 \rfloor$ 
5:  $neighbors \leftarrow \mathbf{0} \in \mathbb{Z}^{n \times k_{\text{neighbors}}}$ 
6: for  $pos \leftarrow 0$  to  $n - 1$  do
7:    $left \leftarrow \max(0, pos - window)$ 
8:    $right \leftarrow \min(n, pos + window + 1)$ 
9:    $window_{\text{idx}} \leftarrow [left, left + 1, \dots, right - 1]$ 
10:  if  $|window_{\text{idx}}| < k_{\text{neighbors}}$  then
11:     $pad \leftarrow k_{\text{neighbors}} - |window_{\text{idx}}|$ 
12:    if  $left = 0$  then
13:       $extra \leftarrow [right, right + 1, \dots, right + pad - 1]$ 
14:    else
15:       $extra \leftarrow [\max(0, left - pad), \dots, left - 1]$ 
16:    end if
17:     $window_{\text{idx}} \leftarrow window_{\text{idx}} \parallel extra$ 
18:  end if
19:  truncate  $window_{\text{idx}}$  to the first  $k_{\text{neighbors}}$  entries
20:  for  $j \leftarrow 0$  to  $k_{\text{neighbors}} - 1$  do
21:     $neighbors[pos, j] \leftarrow idx_{\text{sorted}}[window_{\text{idx}}[j]]$ 
22:  end for
23: end for
24:  $neighbors_{\text{unsorted}} \leftarrow \mathbf{0} \in \mathbb{Z}^{n \times k_{\text{neighbors}}}$ 
25:  $neighbors_{\text{unsorted}}[idx_{\text{sorted}}] \leftarrow neighbors$ 
26: return  $neighbors_{\text{unsorted}}$ 
```

Algorithm 4 GRAVITY OPTIMIZE

Require: Kernel vector $kernel_{\text{values}}$, curvature vector $curvature$, neighbor matrix $neighbors_{\text{idx}}$, iterations $n_{\text{iterations}}$, initial gravity constant G_0 , decay parameter λ , event horizon scale $event_{\text{horizon_scale}}$, computation device $device$, random seed $random_{\text{state}}$

Ensure: Selected gene indices $selected_{\text{idx}}$, optimal gene count k_{opt} , masses M , latent positions X

```
1:  $kc \leftarrow \text{flatten}(kernel_{\text{values}})$  on  $device$ 
2:  $curv \leftarrow \text{flatten}(curvature)$  on  $device$ 
3:  $neighbors \leftarrow neighbors_{\text{idx}}$  on  $device$ 
4:  $n \leftarrow |kc|$ 
5:  $curv_{\text{abs}} \leftarrow |curv|$ 
6:  $M \leftarrow \frac{curv_{\text{abs}} - \min(curv_{\text{abs}})}{\max(curv_{\text{abs}}) - \min(curv_{\text{abs}}) + 10^{-8}}$ 
7:  $M \leftarrow M + 0.05$ 
8: set random seed to  $random_{\text{state}}$ 
9:  $X \leftarrow 0.1 \cdot \text{Randn}(n, 3)$ 
10:  $V \leftarrow \mathbf{0}$  with the same shape as  $X$ 
11: for  $t \leftarrow 0$  to  $n_{\text{iterations}} - 1$  do
12:    $G_t \leftarrow G_0 \cdot e^{-\lambda t / n_{\text{iterations}}}$ 
13:    $X_i \leftarrow X$  expanded along a neighbor axis
14:    $X_j \leftarrow X[neighbors]$ 
15:    $direction \leftarrow X_j - X_i$ 
16:    $dist_{\text{sq}} \leftarrow \sum(direction^2) + 10^{-6}$ 
17:    $d_{kc} \leftarrow |kc[neighbors] - kc| + 10^{-2}$ 
18:    $m_i \leftarrow M$  expanded along a neighbor axis
19:    $m_j \leftarrow M[neighbors]$ 
20:    $force_{\text{mag}} \leftarrow G_t \cdot \frac{m_i \cdot m_j}{d_{kc} \cdot dist_{\text{sq}}}$ 
21:   clamp  $force_{\text{mag}}$  above by 5.0
22:    $F \leftarrow \sum(force_{\text{mag}} \cdot direction)$  over neighbors
23:    $A \leftarrow F/M$  expanded along the coordinate axis
24:    $V \leftarrow 0.85V + 0.15A$ 
25:    $X \leftarrow X + V$ 
26: end for
27:  $center_{\text{of\_mass}} \leftarrow \frac{\sum(X \cdot M)}{\sum M}$ 
28:  $dists \leftarrow \|X - center_{\text{of\_mass}}\|$ 
29:  $top_{\text{mass\_idx}} \leftarrow \text{TopK}(M, \max(10, \lfloor 0.01n \rfloor))$ 
30:  $R_{\text{limit}} \leftarrow \text{Quantile}_{0.90}(dists[top_{\text{mass\_idx}}]) \cdot event_{\text{horizon\_scale}}$ 
31:  $selected_{\text{mask}} \leftarrow (dists \leq R_{\text{limit}})$ 
32:  $selected_{\text{idx}} \leftarrow \text{Where}(selected_{\text{mask}})$ 
33:  $k_{\text{opt}} \leftarrow |selected_{\text{idx}}|$ 
34: return  $selected_{\text{idx}}, k_{\text{opt}}, M, X$ 
```

Appendix B: Parameter Description

Table S1: Core Parameters Used in the proposed method-based feature selection framework

Parameter	Value	Role in the Method
KNN_CURVATURE	200	This parameter defines the neighbourhood size during curvature estimation. For each gene, the variance of distance is calculated within its local neighbourhood to estimate local geometric curvature. A larger value indicates smoother and more globally stable curvature estimation, whereas a smaller value increases local sensitivity.
N_GRAVITY_ITERS	50	This parameter defines the number of iterations used in the gravitational-search-based gene selection method. During each iteration, gene interactions based on simulated gravitational forces in latent space are calculated, leading to progressive refinement of the spatial organization of informative genes. Higher values allow more stable convergence of the optimization dynamics.
GRAVITY_K_NEIGHBORS	70	This parameter determines the number of nearest neighbouring genes participating in the gravitational interaction for each gene. It helps build the interaction graph between genes and controls the local interaction topology of the optimization process. Larger values increase interaction density and global cohesion, while smaller values emphasize localized structural relationships.

Table S2: Classification Experiment Parameters for the Experimental Pipeline

Classifier	Parameter	Value
SVM	Kernel	RBF
	C (Regularization)	10^{-2}
	Gamma	10^{-3}
	Number of Iterations	50
	K-fold Cross Validation	10
XGBoost	Number of Estimators	50
	Learning Rate	0.01
	Booster	<code>gbtree</code>
	Max Depth	5
	Number of Iterations	50
	K-fold Cross Validation	10

Table S3: Differential Expression Analysis Parameters for DESeq2, Voom-Limma, and NOISeq Pipelines

Method	Parameter	Value / Role
Voom-Limma	Normalization Method	TMM normalization using <code>edgeR::calcNormFactors()</code>
	DEG Significance Threshold	Benjamini-Hochberg adjusted p-value < 0.05
	Fold Change Threshold	$ \log_2 \text{FC} > 1.0$
	Statistical Model	Linear modelling using <code>lmFit()</code> with empirical Bayes moderation using <code>eBayes()</code>
	DEG Extraction Strategy	Pairwise contrasts for multi-class datasets followed by union of significant DEGs
	Transformation	Voom $\log_2$ -CPM precision-weighted transformation
NOISeq	Probability Threshold (q)	0.8 for DEG identification
	Normalization Method	TMM normalization using <code>norm = "tmm"</code>
	Number of Iterations (r)	20
	CPM Filtering Threshold	CPM ≥ 1 in at least one group
	Pseudocount Parameter (k)	0.5
	Adjustment Parameter (adj)	1.5
	Random Seed	42
	DEG Extraction Strategy	Pairwise comparisons across classes followed by union of DEGs
DESeq2	Normalization Method	Median-of-ratios normalization
	DEG Significance Threshold	Benjamini-Hochberg adjusted p-value < 0.05
	Fold Change Threshold	$ \log_2 \text{FC} > 1.0$
	Dispersion Estimation	Negative binomial dispersion modelling
	Statistical Test	Wald significance test
	DEG Extraction Strategy	Pairwise differential analysis with union of DEGs for multi-class datasets
	Random Seed	42

Appendix C: Gene Set Enrichment and KEGG Pathway Analysis Figures

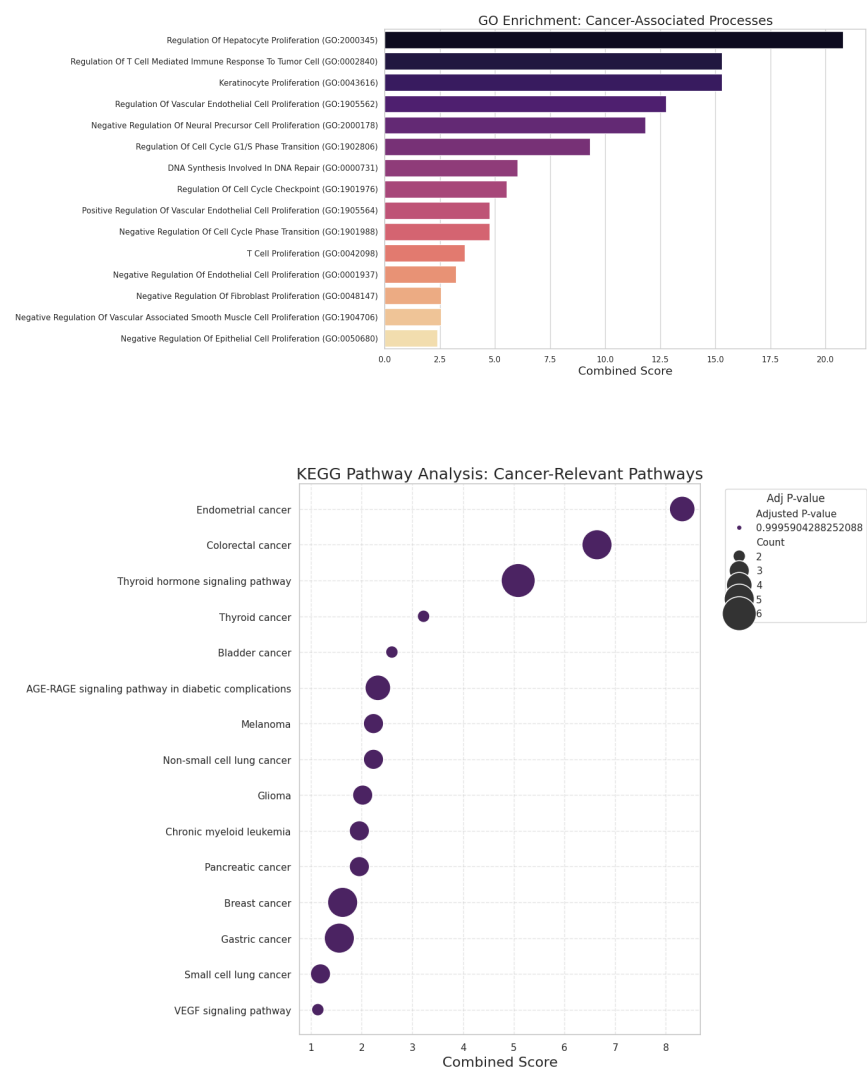


Figure S1: Gene set enrichment analysis and KEGG pathway enrichment analysis for DEGs selected from TCGA LUAD

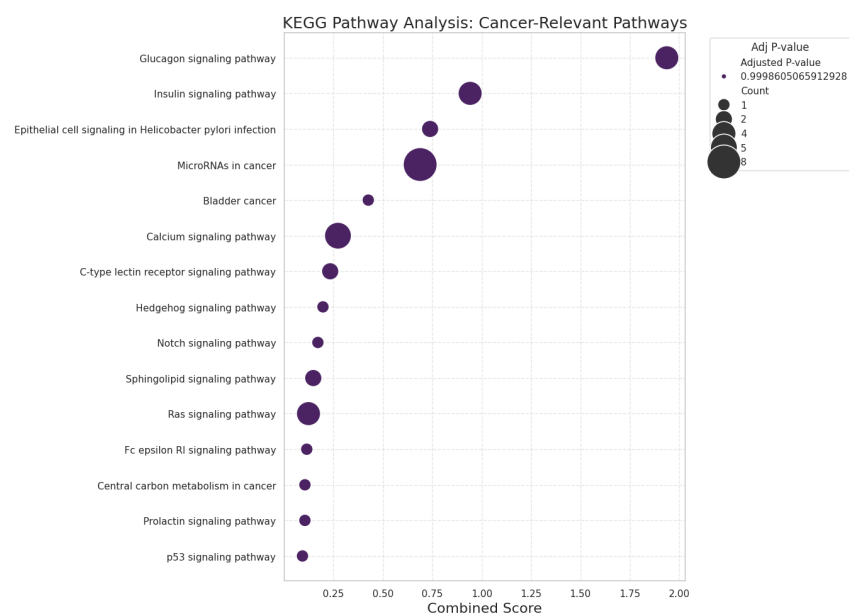
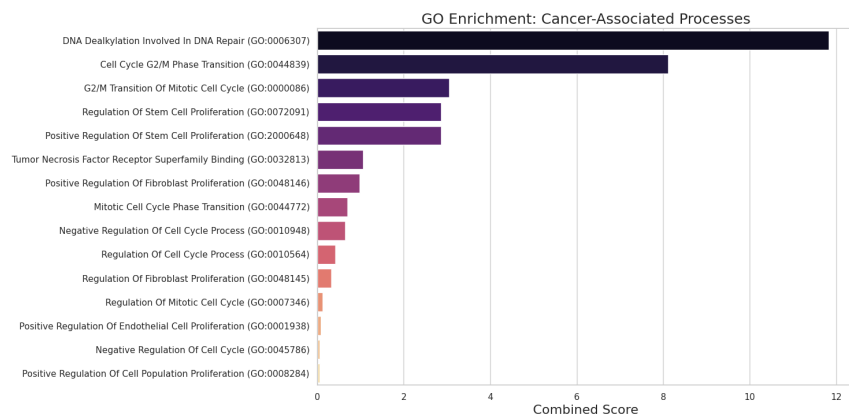


Figure S2: Gene set enrichment analysis and KEGG pathway enrichment analysis for DEGs selected from TCGA HNSC

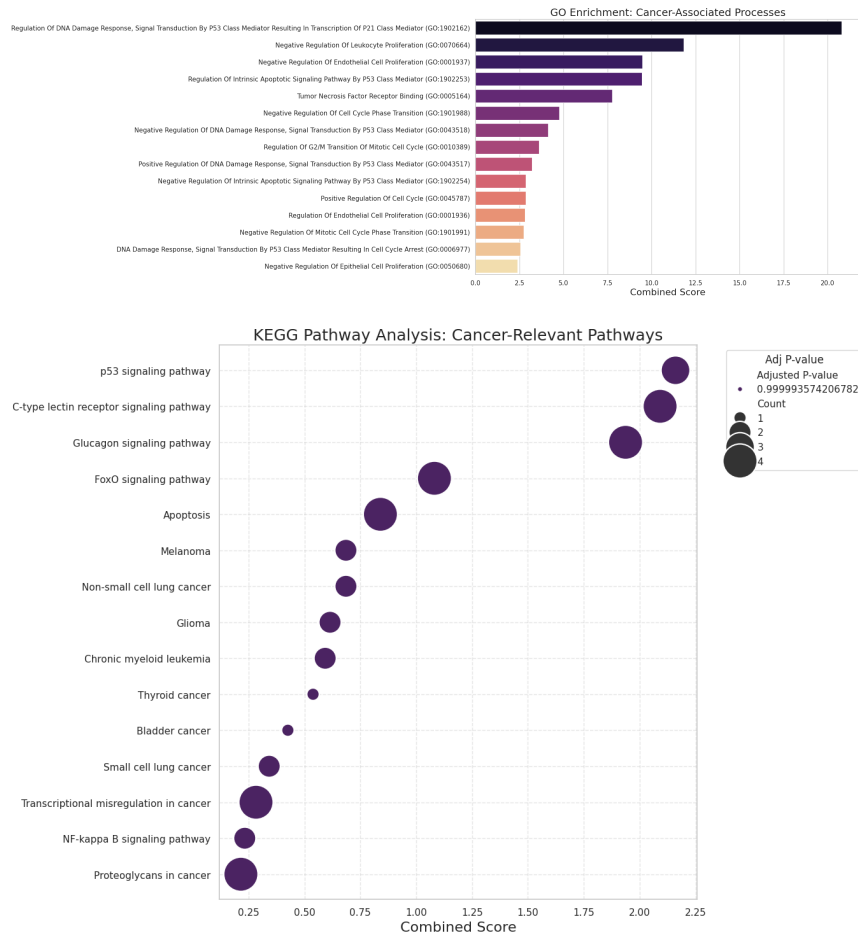


Figure S3: Gene set enrichment analysis and KEGG pathway enrichment analysis for DEGs selected from TCGA BRCA

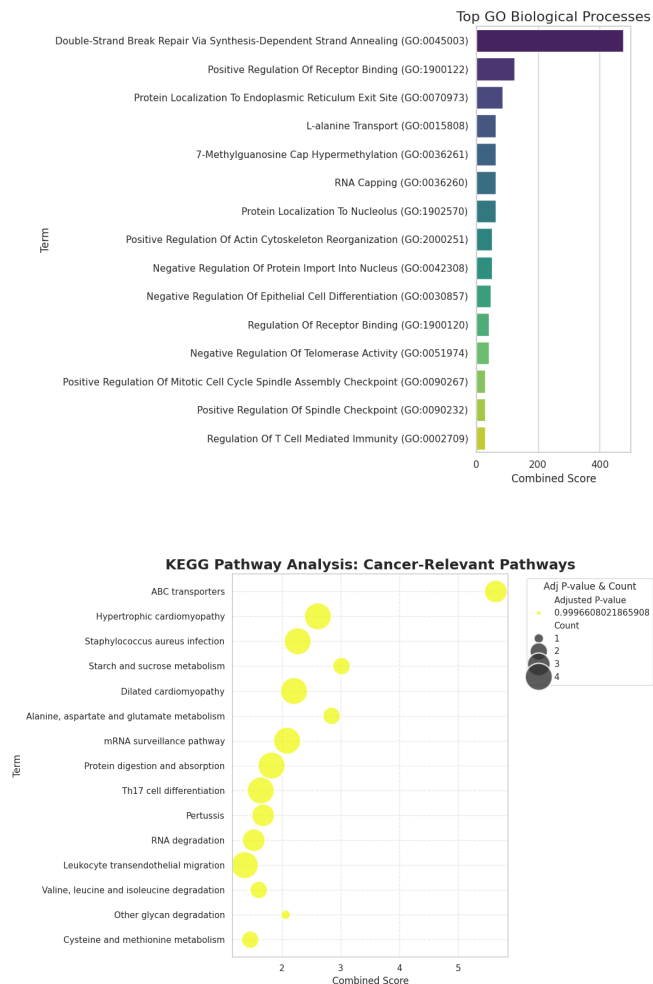


Figure S4: Gene set enrichment analysis and KEGG pathway enrichment analysis for DEGs selected from Gene Expression Omnibus Liver Cancer

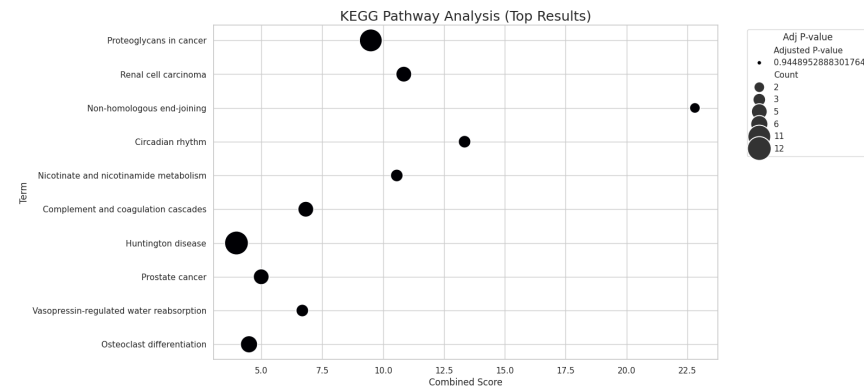
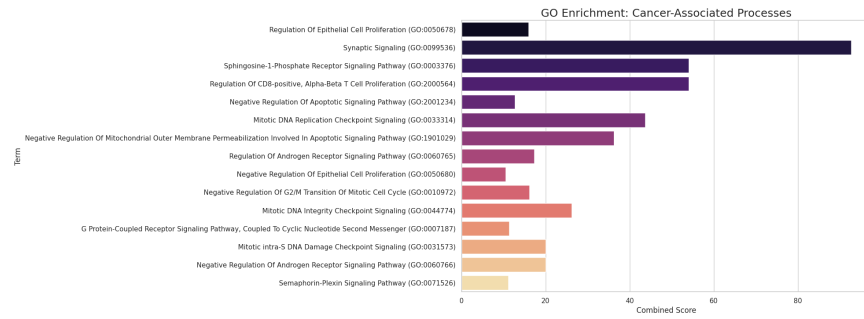


Figure S5: Gene set enrichment analysis and KEGG pathway enrichment analysis for DEGs selected from Gene Expression Omnibus Throat Cancer



Figure S6: Gene set enrichment analysis and KEGG pathway enrichment analysis for DEGs selected from Gene Expression Omnibus Pancreatic Cancer